Towards Seamless Integration in a Multi-modal Interface

Dennis Perzanowski, William Adams, Alan C. Schultz and Elaine Marsh

Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Washington, DC 20375-5337
<dennisp | adams | schultz | marsh>@aic.nrl.navy.mil
http:// www.aic.nrl.navy.mil/ <~dennisp | ~adams | ~schultz | ~marsh>
voice phone: 202-767-9005
fax: 202-767-3172

Abstract

We are designing and implementing a multi-modal interface to an autonomous robot. For this interface, we have elected to use natural language and gesture. Gestures can be either natural gestures perceived by a vision system installed on the robot, or they can be made by using a stylus on a Personal Digital Assistant. In this paper we describe how we are attempting to provide a seamless integration of the various modes of input to provide a multi-modal interface that humans can manipulate as they desire. The interface will allow the user to choose whatever mode or combination of modes seems appropriate for interactions with the robot. The human user, therefore, does not have to be limited to any one mode of interaction, but can freely choose whatever mode is most comfortable or natural.

Introduction

We are investigating human-computer interaction in the context of communicating with an autonomous robot. Natural language is one means of human-computer interaction. However, a problem immediately arises when one considers natural language.

In natural language certain linguistic elements are ambiguous, unless they are accompanied by additional information, such as when someone says “Go over there.” Unless this utterance is accompanied by a gesture or a nod of the head in a particular direction--some overt cue, the intended location is ambiguous. Interacting with a computer or a robot that can understand natural language can be difficult unless the designers of the interface incorporate more than one means of interaction to mitigate this built-in ambiguity in natural language. In the past, interface designers have incorporated mouse operations, touch-screens, and various other multi-modal devices to assist users of these systems to handle so-called *deictic* elements, such as "this," "that," "there," "that door," and "that door over there."

While there may be several means for disambiguating deictic elements in speech, we do not consider the purely linguistic solutions, such as clarification or repetition. We have focused on another communication technique for disambiguating deictic elements in speech; namely, using natural and synthetic gestures, which we will define shortly. Since humans frequently use gestures during their communicative acts, we have incorporated gestural communication in our interface to an autonomous robot.

The interface can disambiguate deictic elements in the speech output by means of natural gestures obtained using a rangefinder mounted on the top of the robot. We are expanding this interface to incorporate pointing and drawing on a PDA, a Personal Digital Assistant, display. For this work we have been using a hand-held Palm V Organizer. It is being used in command and control situations, along with natural gestures and natural language, to direct the robot, and to disambiguate deictic elements in commands. The freedom to use any combination of input devices in this interface is the focus of this paper.

In our initial research, we focussed on natural language and natural gestures in command and control situations to an autonomous robot. However, our interest has now expanded to incorporate a wider variety of communicative ways to disambiguate deictic elements. Specifically, we are interested in combining a mechanical medium of interaction with natural language and natural gesture to allow people to switch modes of interaction conveniently when communicating with a robot. We are interested in creating a seamless interface which allows users to move back and forth freely between natural and what we are calling *synthetic* interactions. This will require a robotic system robust enough to know the difference.

To understand how we incorporate these different modalities in our system and are attempting to design a system with the robust capabilities outlined above, we must first briefly discuss how we categorize gestures.

Categorizing Types of Gestures

While some gestures in human communication are redundant, such as coincidental movement of one's hands

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2000		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Towards Seamless Integration in a Multi-modal Interface	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Navy Center for Applied Research in Artificial Intelligence,Naval Research Laboratory,Code 5510,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

We are designing and implementing a multi-modal interface to an autonomous robot. For this interface, we have elected to use natural language and gesture. Gestures can be either natural gestures perceived by a vision system installed on the robot, or they can be made by using a stylus on a Personal Digital Assistant. In this paper we describe how we are attempting to provide a seamless integration of the various modes of input to provide a multi-modal interface that humans can manipulate as they desire. The interface will allow the user to choose whatever mode or combination of modes seems appropriate for interactions with the robot. The human user, therefore, does not have to be limited to any one mode of interaction, but can freely choose whatever mode is most comfortable or natural.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**7**	

while speaking, many other gestures accompanying human communication provide information about the content of what is being said. Some gestures provide information about the user's emotional state, or emphasize aspects of the speaker's utterance, such as highlighting or underlining in text. For example, excited hand waving or repeatedly pointing one's finger into the palm of one's hand while speaking might indicate the speaker's excitement or even how seriously and important the speaker considers the point that is being made. For our research, however, we are only concerned with gestures that provide a particular kind of information; namely deictic information, that is, clarifying information about the location of some region in the world or some object referred to. For our purposes we categorize these deictic or pointing gestures as either natural or synthetic. We turn now to this distinction.

Swinging one's arm, or pointing one's finger, in a particular direction to indicate some location or an object in the environment is natural. People point to locations and objects in the real world frequently while they are speaking, and the referents or objects referred to are immediately available in the environment. Gestures in these contexts assist individuals communicating with each other to know exactly what location or object the speaker is referring to and may even crucially disambiguate what location or object is being referred to.

Thus, a command, such as "Go over there," can be disambiguated with an accompanying natural gesture on the speaker's part. In such a case, the speaker might point with his/her finger in a particular direction to indicate the location being referred to in the utterance. Viewers of such a gesture interpret both the speech signal and gesture, understanding that the gesture disambiguates a specific element in the speech signal.

A synthetic gesture is very similar to a natural one. Instead of using some natural way of communicating a gesture, such as swinging an arm in a particular direction or pointing a finger to a location or object, a synthetic gesture is made by utilizing some mechanical device to convey the information, such as a mouse click or the use of a finger or stylus on a touch-screen. It should also be noted here, that we are using the term *synthetic* somewhat differently from the kinds of *synthetic* gestures employed by other systems, such as (Kortenkamp, Huber, and Bonasso 1996). Gestures for them are arbitrary in the sense that the gesture has no inherent meaning; it is merely being used as a symbol that maps to some action.

For us, synthetic gestures are those made by pointing to some representation of real objects and/or locations. For example, a user says "Go over there" and points to an icon representing a door on a display. Since the icon is a pictorial representation of a real object, we categorize this action as a synthetic gesture.

Natural Language Processing

Since one mode of interaction in our interface is verbal, we turn now to a brief discussion of the natural language processing in the interface. A more detailed description of how we process the natural language input is discussed elsewhere (Perzanowski et al. 1998).

In our interface, users can speak into a wireless microphone. The auditory signal is then processed by a domain-specific grammar and dictionary that we have developed using IBM's ViaVoice speech recognition system. The signal is translated into a text string, which is then parsed and interpreted by our in-house natural language understanding system, Nautilus (Wauchope 1994), producing a logical form. This resulting representation is correlated with the gesture data obtained from the robot described in the next section.

Nautilus employs a robust natural language parser which derives a complete syntactic analysis and a logical representation of the input string. We believe we require a richer representation of the speech input for discourse processing than stochastically based parsers (Tomita 1986) can provide. We need detailed syntactic and semantic representation to map gestures, commands, and their associated goals. We cannot rely solely on a probabilistic technique of acquiring a linguistic understanding of the input, for during extended dialogs with the robot, fragmentary input may occur, such as in the short dialog (1-3) where certain linguistic elements of the dialog are missing.

(1) Go over there.
(2) Where?
(3) Over there.

Having a complete representation of the sentences of the dialog, not just a partial and probabilistic record of the syntactic elements present in the speech signal, allows us to process discourse elements later without having to re-parse the sentences because elements are missing. Therefore, our grammar fills in missing elements, such as "go" in (3), based upon the surrounding dialog. Later processing is not held up, so to speak, because some sentences of the dialog are missing elements.

Similarly, gestures can be somewhat vague in terms of what specific object or location is being referred to. We, therefore, opt for as much detailed information from all our input sources as is possible. Furthermore, we have recently begun to track goal information and whether or not goals have been attained (Perzanowski et al. 1999). A full syntactic and semantic parse of each utterance provides us with a high level of confidence for tracking goals and whether or not they have been obtained.

In our system, linguistic information is coupled with gestural information, whether it is input from a rangefinder on the robot or obtained from a touch-screen on a PDA. Since gestures can come from one of two sources, we turn now to how the natural gestures obtained

from a rangefinder or synthetic ones obtained from interactions with a PDA are interpreted.

Gesture Processing

Although we are using several robots, Nomad 200s, XR-4000s, and an RWI ATRV-Jr, we will limit our discussion here to the Nomad 200, manufactured by Nomadic Technologies, Inc.

A process running on the robot is used to interpret natural gestures given by the human user. The gestures are detected using a structured light rangefinder which emits a horizontal plane of laser light 30 inches above the floor. A camera fitted with a filter tuned to the laser wavelength is mounted on its side. Given that the laser and camera mount are at a right angle, and the camera is tilted a fixed amount, the distance to a laser-illuminated point can be easily triangulated. With the sensor, the robot is capable of tracking the user's hands and interpreting their motion as vectors or measured distances. The method of obtaining this type of gesture is found in (Perzanowski et al. 1998).

However, suffice it to say here, once a gesture is perceived, it is identified as either a continuous or a stationary gesture. Gestures are queued and the most recent gesture from the queue is checked for appropriateness and validity for a particular command.

In addition to the perception of natural gestures in the real world, we have incorporated the stylus-based touch-screen capability afforded by the PDA. It can be used to provide synthetic gestures indicating location, direction, and other deictic information. In order to show how the PDA provides these synthetic gestures, we'll first briefly describe the underlying system which this interface is controlling.

(Schultz, Adams, and Yamauchi 1999) presented an integrated system of robot exploration, mapping, localization, and adaptation. In summary, the robot can explore a new environment and produce a map. The robot localizes against the map and updates the map to reflect changes in the room. Given a goal location, this map is used by a path planner and a reactive navigation algorithm to move the robot along the shortest path to the goal while avoiding obstacles not appearing in the map. Persistent obstacles are quickly incorporated into the map and the path planner will re-plan around them.

Figure 1 shows a Palm V Organizer running our interface software. The central feature is the map of the environment, which is regularly updated to reflect changes made by the robot in response to sensor readings. The robot's current location is marked by an individualized letter, 'R,' on the map and its location is regularly updated. Individual robots can be addressed by selecting them from the menu of buttons along the top of the display. Information is transmitted between the PDA and the rest of the system via TCP/IP by means of a Mercury wireless ethernet device connected to the PDA's serial port. The device sees the connection, and the Mercury device provides the host and port for the TCP/IP connection from outside machines.

Figure 1: Palm V Organizer, stylus, and Mercury ethernet device

Currently, two types of synthetic gestures can be entered on the PDA. The first type is a single location, indicated by tapping at a location on the map. If appropriate for the command, the location can also be mapped to a specific object at that location. The second type is a general area, selected by dragging the stylus across the map.

Combining Commands and Gesture Understanding

In this section, we will discuss the overall architecture of the multimodal interface, as it is represented in Figure 2.

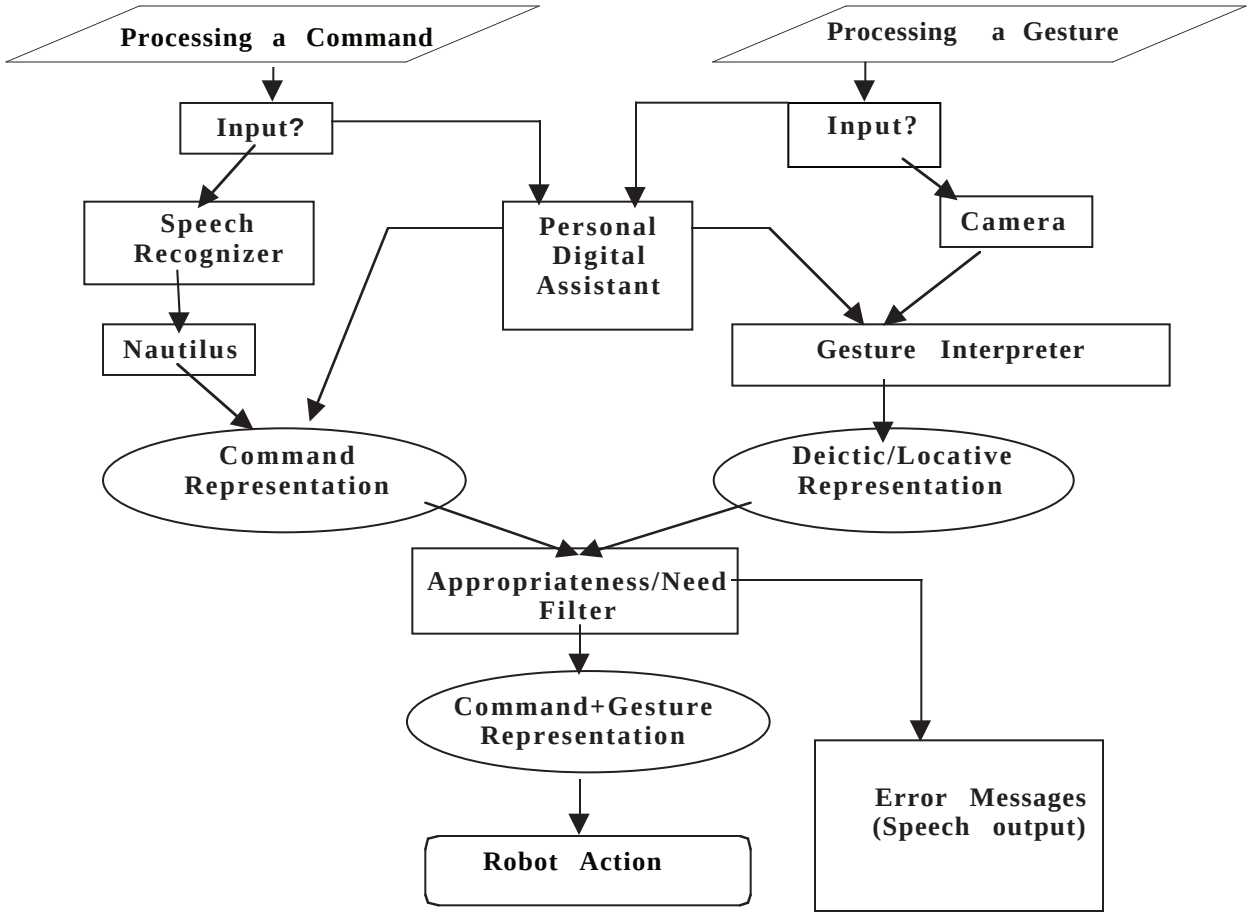

Figure 2: Schematic Overview of Multi-modal Interface

If a command is spoken, the acoustic signal is mapped to a linguistic string that is parsed and interpreted by Nautilus. The spoken command is translated into a logical representation of the spoken utterance. (4) is an example of this mapping.

(4) Go to the waypoint over there →
(imper (:verb gesture-go
(:agent (:system you))
(:to-loc
(waypoint-gesture))
(:goal (there))))

Since we now track goals and whether or not they have been achieved, we have incorporated a placeholder for this information, as in (5), where the last element of the list simply notes whether or not the goal has been completed.

(5) (imper (:verb gesture-go
(:agent (:system you))
(:to-loc
(waypoint-gesture))
(:goal (there)))) 0)

As goals are achieved, this placeholder is updated. This representation is mapped to specific robot commands in one of Nautilus' translation components so that the robot can process this data.

However, natural speech is not the only means of providing commands. The PDA can also be used to input commands by tapping any of a number of menu buttons seen along the right side of the display.

The gesture is resolved against the command, be it from natural language or a tap on the menu button of the PDA, and the appropriate robotic action is taken or a spoken error message is produced, alerting the user to the kind of error that was produced and perhaps asking for further clarification. The mapping of the command and gesture input to a response is a function of the presence or

absence of a gesture accompanying the speech input and the appropriateness or inappropriateness of that gesture for the given utterance.

At present, for example, users can touch the "Follow" button in order to tell the robot to follow a path, and then draw a path on the display. The robot will follow the path, avoiding any unexpected obstacles along the way.

Since we have incorporated several modalities for inputting both commands and gestures, we allow the user to decide which means of interaction he or she wishes to employ to communicate a command, deictic or locative information to the robot. Thus, the user may either speak a command or tap a corresponding button on the PDA's screen. If the user points to objects or locations in the real world, the robot "sees" the gesture through its structured light rangefinder. If the user points to locations and/or objects on the screen, the pixel information on the screen is translated to objects and locations in the real world.

The user is free to use any combination of input modalities. For example, the user may say "Go over there" or click on a menu choice "Go to"--as indicated by one of the buttons on the right-hand side of the PDA's display--and simultaneously either gesture naturally by sweeping his/her arm in some direction in the real world, or point to a location on the PDA's map.

When insufficient information is provided, the system will appropriately complain and ask the user for additional information of whatever type was lacking.

The burden of obtaining and interpreting information, therefore, is on the system, not upon the user to determine what means of interaction is appropriate. The system is aware of which gestural source--natural gesture or stylus pointing—and which command source--natural speech or button menu--is acting as input. While the user will ultimately be free to interact with any of the modalities of the interface, crossing modalities is currently somewhat limited: it does not accept natural gestures along with button menu items. We are currently designing and modifying the system to reflect the two parallel capabilities of gesture and command understanding, as they are represented in Figure 2, and to permit greater freedom of interaction.

As we have portrayed it here, our system is divided into two parts, one part processes the commands, and the other processes gestures. Both parts, however, must determine at first where the input for its components is coming from. In the case of a command, the system must determine if the command is being given verbally or by a click of a button on the PDA. When a command is obtained, either Nautilus or the Gesture Interpreter must interpret the input and translate it into an appropriate representation suitable for the robotic system to handle.

On the gestural side of the system, gestural input can be obtained either visually from the rangefinder mounted on the top of the robot, or by user interaction with the touch-screen on the PDA. The gesture must be interpreted and transformed into an appropriate representation. This representation and the command representation are then checked in a filtering component that ensures that an appropriate gesture was perceived with a particular command, or if an inappropriate gesture was obtained, an error message is produced. Likewise, if some kind of gesture was required to resolve ambiguity, the system must so inform the user. For example, a gesture accompanying the utterance "Go to that table" only makes sense when that gesture is made to an actual table in the environment. Pointing to nothing or to a chair while uttering the above sentence is either meaningless or confusing, and the system must so inform the human user so that subsequent corrective action can be taken. We provide feedback to the user by means of spoken error messages that request specific additional information or clarification. Thus, for example, the system would not respond in the above example simply by saying "I don't understand," but would offer the user more help by saying "You told me to go to a table, but you pointed to a chair. I'm confused. What do you want me to do?"

With a system that tries to figure out what the means of interaction are and when errors occur what corrective information is needed, the user is freed to interact with it as he or she sees fit.

An Example

At this point, we would like to provide a concrete example of the capabilities of our multi-modal interface to an autonomous robot.

As we have said above, commands to the robot can come from either of two sources: either the user can verbally direct the robot to perform some action, or the user can click a button on the handheld PDA. A command like "Go over there" might be spoken, or a button on the PDA display labeled GO TO might be pushed. In the first case, the verbal command, our natural language processing modules analyze the verbal input by transforming the auditory input into a text string, which is then parsed and translated into an appropriate robot command. If the user pushes the GO TO button on the PDA display, the output is mapped to the same linguistic string as mentioned above and the processing continues to obtain a robot command. Therefore, it does not matter whether the robot command is initiated verbally or by clicking a button on the PDA display. Either input is translated into a linguistic string for interpretation.

However, the utterance "Go over there" without some kind of gesture is meaningless. To provide the system with the necessary locative information, the user might swing his/her arm in a particular direction and point, and the rangefinder on the robot calculates where the desired location is. On the other hand, the user might click on some set of coordinates on the PDA display map. With a set of PDA coordinates, commands and gestural information are combined, and the information is translated into an appropriate robot command.

If the user doesn't gesture or click on an appropriate place on the PDA display, the robot responds with an informative message. In this case, the robot might say

something like “You told me to go somewhere, but you didn’t tell me where.” The user then is free to either gesture naturally or use the PDA to supply the appropriate information. Fragmentary verbal input, such as simply saying "Over there" at this point in the interaction is acceptable, since the system keeps track of the dialog, as we have discussed earlier.

We have designed the system so that the user does not have to concentrate on which mode he/she wishes to use. Inputs can come for either of the two command input sources, and they can be matched with any of the other gestural input modalities of the interface. Thus, the user is not restricted or confined to matching modalities.

Future Work

While we are very much interested in developing a seamless integration of modalities in this interface, we are also interested in expanding the kinds of deictic and location information that the system can process. Incorporating a PDA into the interface allows us to interact in a more complicated way with the robot than we have previously.

For example, a user might want to ask the robot, "Is there a door in the room?" Utilizing its vision system, the robot could then explore the area to determine the answer to the question. Or, if the user thought there is a door in the area, but was unsure where, the user could ask the robot, "Is there a door over there?" and point to some location in the environment. Further, the user might ask the same question, but if the location of the door was more uncertain, the user might sweep his/her hand or arm in some generally indicated direction in the room. Similar gestural actions could accompany these queries while the user employed a PDA with a touch-screen.

In either case, however, the gesture accompanying the linguistic term "there" or "over there" is not interpreted as a single point, but as some more generalized area close to where the user pointed or swept his/her arm or stylus.

It would, therefore, be incumbent on the system to know that even a single location indicated by x,y coordinates on a PDA’s screen might not be indicative of a single location in the real world, but might indicate a more generalized area. In any event, we need to explore a wider and richer range of gestures, whether natural or synthetic.

Likewise the semantics of locative gestures becomes even more complicated in such a situation as when someone might say, "Explore this area" and point in a general direction in the real world or to a single location on a touch-screen. The speaker might not mean for the system to explore one particular limited spot in the real world or by a single spot as indicated by the x,y coordinates pointed to on a touch-screen. The system would have to be smart enough to know that such an utterance with either gesture meant for the system to explore a particular region of some arbitrary distance or radius.

However, if the speaker were to say, "Explore this area" and lasso a specific area on a touch-screen, clearly the intended location would be more limited. But the situation is still complex if the utterance is accompanied by a sweeping gesture in the real world. A single movement of the arm, for example, might be indicative of a predefined area to be explored, while a sweeping motion of the arm, which is repeated, might indicate a more general area for exploration. The interpretation of gestures and the semantics of deictics is quite complex and requires further investigation. We are currently working on this group of related problems and incorporating them into the system.

We are also designing an experiment in which human subjects command an autonomous robot to perform a task, using the interface we have already constructed. Our purpose in doing so is twofold: first, we hope to determine in what ways our interface fails when humans use it. This will improve what we have already accomplished. Second, we hope that such an experiment will point out ways humans interact with robots, and given the kinds of interaction we currently permit, what our system's limitations are and where it needs to be expanded. Our work here will be conducted in conjunction with human factors and cognitive scientists.

Finally, another kind of gesture that we currently do not address is an unambiguous deictic gesture that is unaccompanied by an explicit command. For example, someone might perform a beckoning motion with his/her hand or finger without saying "Come over here," and the observer might still interpret the gesture correctly. It alone is sufficient to convey the meaning. We currently do not address gestures made in silence. We hope to address these issues in our continuing research on our multi-modal interface.

Conclusions

We believe the interpretation of various kinds of gestures and the various modes chosen for inputting that information is a much more complex situation than we have presented here. However, in a very limited sense, we are disambiguating the possible referents for locative and deictic information in a multi-modal interface to an autonomous robot, and we are trying to provide a seamless integration of the modes of interaction in the interface to allow the user greater freedom and flexibility when attempting to interact with that system.

This interface utilizes spoken utterances and gestures that are either executed in the real world and perceived by a rangefinder mounted on the top of a mobile robot or are interpreted from various interactions with a stylus and a PDA held by the user.

The user can command the robot to perform certain actions and accompany these commands with a variety of gestures. The system disambiguates locations which may

be inherent either in the spoken command or in the location indicated. We have used principled techniques for disambiguating speech: deictic elements in natural language require some sort of referent--either linguistic or gestural--to disambiguate them. We are currently working on allowing the user to freely interact with the system. Thus various combinations for inputting a command--speech vs. menu button--and indicating locative or deictic information--natural or synthetic gesture--are possible. The user need not be concerned with how to disambiguate information. It is incumbent on the system. The user can simply communicate information in whatever mode seems most efficient and/or natural.

Acknowledgments

This work is funded in part by the Naval Research Laboratory and the Office of Naval Research.

References

Kortenkamp, D; Huber, E.; and Bonasso, R.P. 1996. Recognizing and Interpreting Gestures on a Mobile Robot. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, 915-921. Menlo Park, CA: AAAI Press.

Perzanowski, D.; Schultz, A.C.; Adams, W. 1998. Integrating Natural Language and Gesture in a Robotics Domain, " In *Proceedings of the IEEE International Symposium on Intelligent Control: ISIC/CIRA/ISIS Joint Conference*, 247-252. Gaithersburg, MD: IEEE Press.

Perzanowski, D.; Schultz, A.C.; Adams, W.; and Marsh, E. 1999. Goal Tracking in a Natural Language Interface: Towards Achieving Adjustable Autonomy. In *1999 IEEE International Symposium on Computational Intelligence in Robotics and Automation: CIRA '99*, 208-213. Monterey, CA: IEEE Press.

Schultz, A.; Adams, W.; and Yamauchi, B. 1999. Integrating Exploration, Localization, Navigation and Planning With a Common Representation. *Autonomous Robots 6(3):* 293-308.

Tomita, M. 1986. *Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems.* New York, NY: Kluwer Academic Publishers.

Wauchope, K. 1994. Eucalyptus: Integrating Natural Language Input with a Graphical User Interface, NRL Technical Report, NRL/FR/5510-94-9711, Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC.